Topics in Data Visualization

Describing Graphics

Jun 27 2014

Charlotte Wickham

vis.cwick.co.nz

We are moving…

From Monday we will meet in ALS 4000

While you wait....

Draw a scatterplot of goals versus games

by hand!

Player	Team	Goals	Games
Lionel Messi	Argentina	4	3
Luis Suárez	Uruguay	2	2
Thomas Mueller	Germany	4	3
Robin Van Persie	Netherlands	2	3

I sent an email to everyone I have on my list Wednesday.

If you didn’t get it email me!

From last time…

Keep things simple:

Be judicious with ink

Don’t use extraneous dimensions

Tell the truth:

The representation of numbers should be proportional to the numbers themselves.

Go easy on the viewer:

Clear labeling

Avoid full saturation colors

From last time…

ggplot2

By default:

- subtle grid
- forces you to use a coordinate system

You have to work very hard to:

- break axes/coordinate systems
- Get 3D effects

Sometimes the answer to “How do I do … in ggplot2” is “You shouldn’t”

From last time…

Places to find bad graphics

The archives at:

http://junkcharts.typepad.com/

Google image search:

bad charts/plots/graphics/infographics

From last time…

Reading

Chapter 1 & 2

The Visual Display of Quantitative Information, Edward Tufte

(available on 3 hour reserve in the library, borrow my copy for an hour or two, or buy it (~$26) and read the whole thing)

Wainer, H. ***How to Display Data Badly*** The American Statistician, Volume 38, Issue 2, 1984

http://www.jstor.org/stable/2683253

Some homeworks already posted!

Today

Describing a graphic in terms of **mappings** between data **variables** and the **aesthetics** of **geometric objects**.

While you wait....

Draw a scatterplot of goals versus games

by hand!

Player	Team	Games	Goals
Lionel Messi	Argentina	3	4
Luis Suárez	Uruguay	2	2
Thomas Mueller	Germany	3	4
Robin Van Persie	Netherlands	3	2

scatterplot - points where horizontal position (x) corresponds to one variable and vertical position (y) to another.

points - a **geometric object**

horizontal position, x
vertical position, y

} **aesthetics** = visual properties of the object

We want to use a point for each player where we:

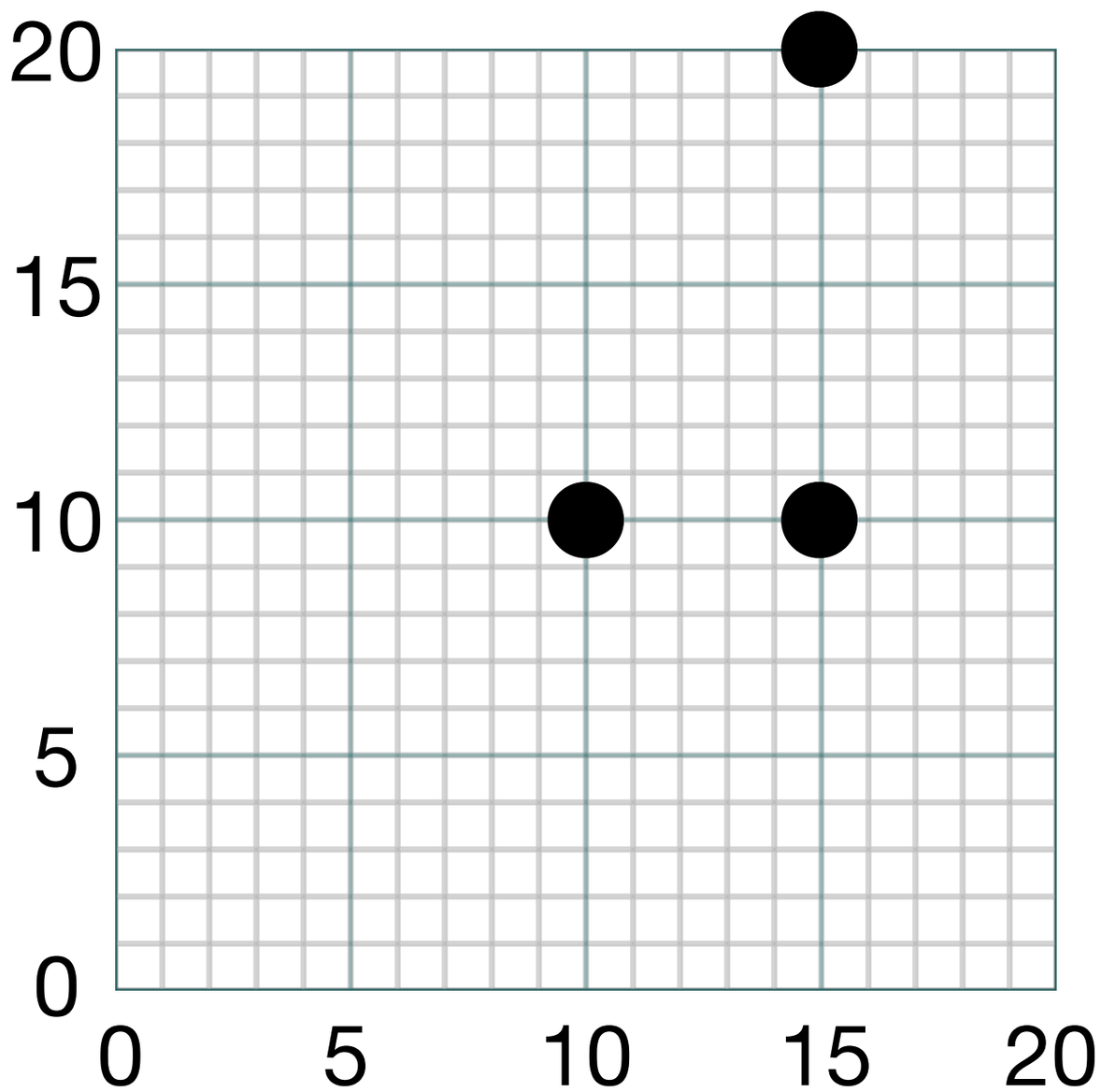
map the data variable Games to the horizontal position of the point

map the data variable Goals to the vertical position of the point

X	Y
3	4
2	2
3	4
3	2

map: we want a direct relationship between the data value and the physical value of the aesthetic in the graphic

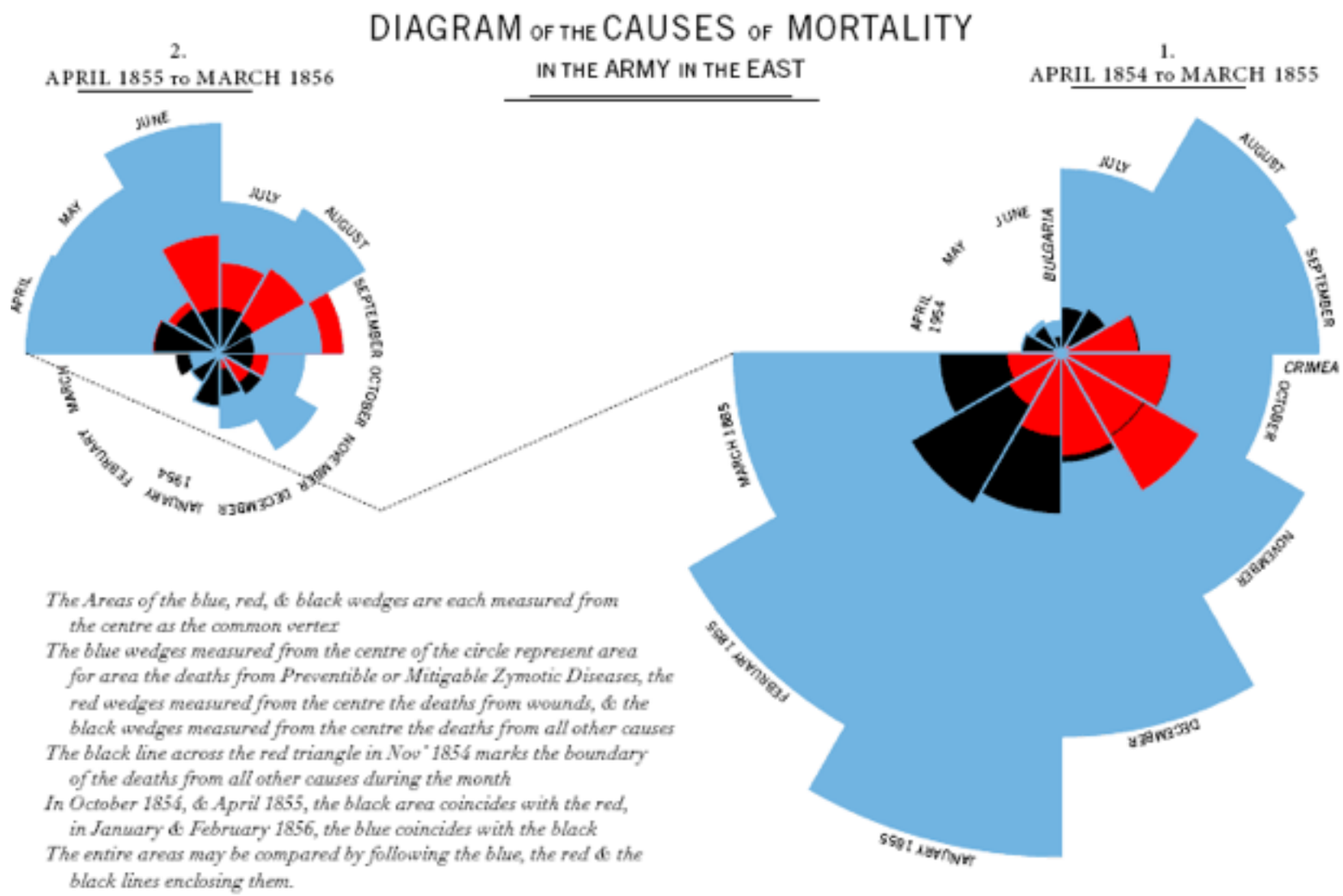
To actually render the plot…



We also need:

- a coordinate system
- scales for each aesthetic

Does this plot have a name?



we don’t want to memorize lots of plot names

we don’t want to be constrained to named plots

Instead we'll describe plots

Describe the data.
(observations? what are the variables?)

Describe how the variables are mapped to aesthetics of the geometric object.

Describe anything unusual about the coordinate system or scales.

Describe any non-data annotations.

perhaps the simplest geometric object

Points

Aesthetics:

x,

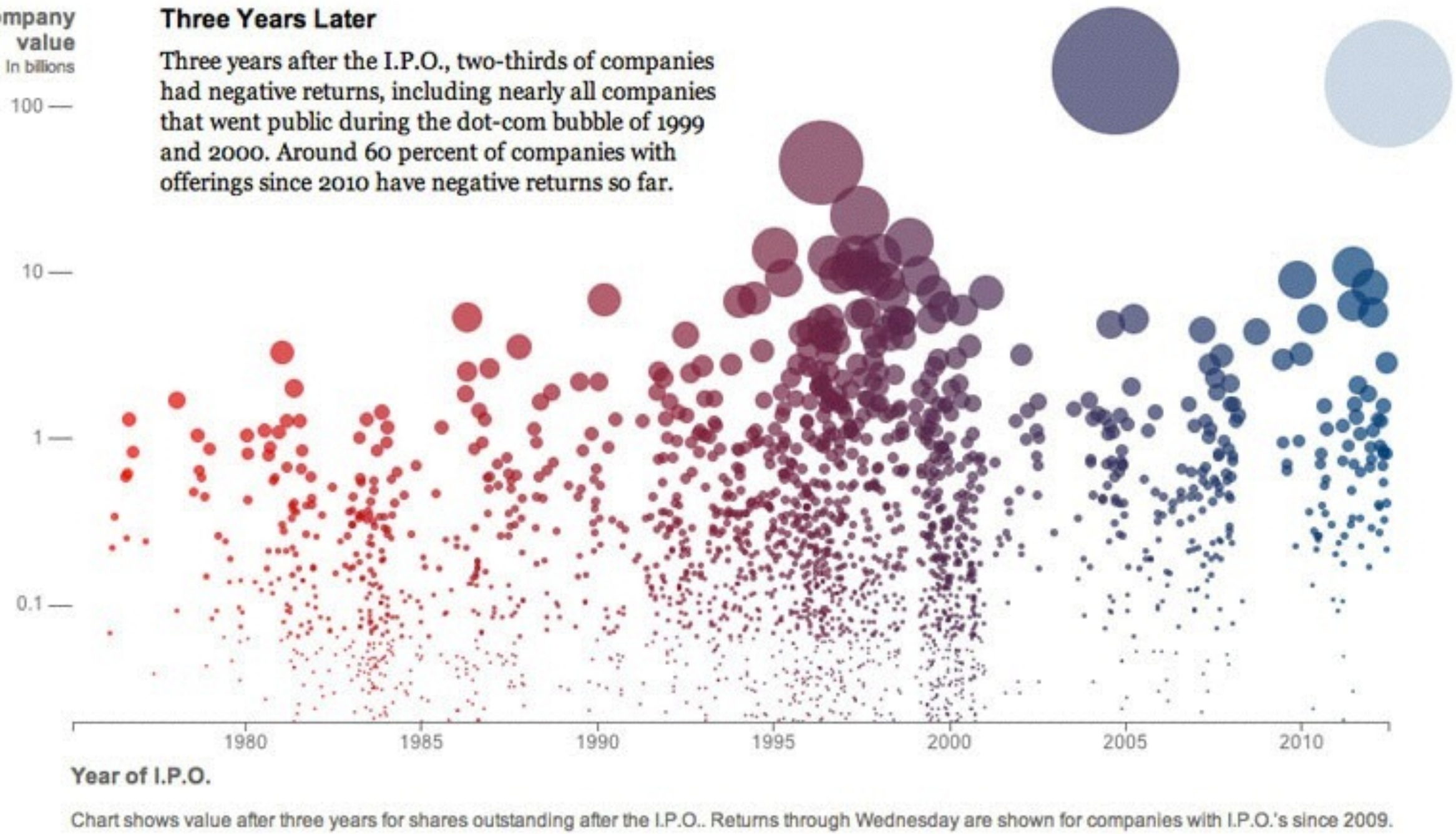
y,

shape, ● ■ ▶

size, • ● ●

colour, ● ● ●

alpha, (transparency) ●●●



http://www.nytimes.com/interactive/2012/05/17/business/dealbook/how-the-facebook-offering-compares.html

The data explore tech companies that have had an I.P.O. since 1980. For each company, we have the date of their I.P.O., and the valuation of that company three years later.

Each company is represented by a (semi-transparent) point. The date of their I.P.O. is mapped to horizontal position and the color of the point.

The valuation of the company three years after their I.P.O. is mapped to the vertical position and the size of the point.

The vertical axis is on a logarithmic scale.

Where the Heat and the Thunder Hit Their Shots

The shooting patterns for the players on the Miami Heat and the Oklahoma City Thunder reveal where they are most dangerous on the court. Below, compare each player's strengths using court maps and analysis by Kirk Goldsberry, a geography professor at Michigan State. Related Article »

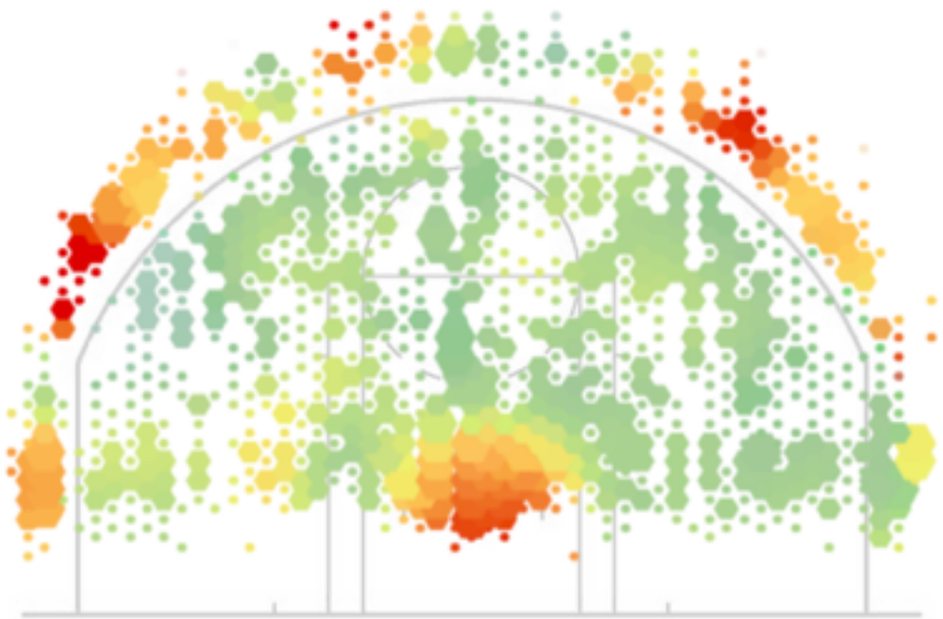
All Shots | 3-Pointers | Midrange | Close Range

Number of attempts: Low – High

Points per region: Low – High

Miami Heat

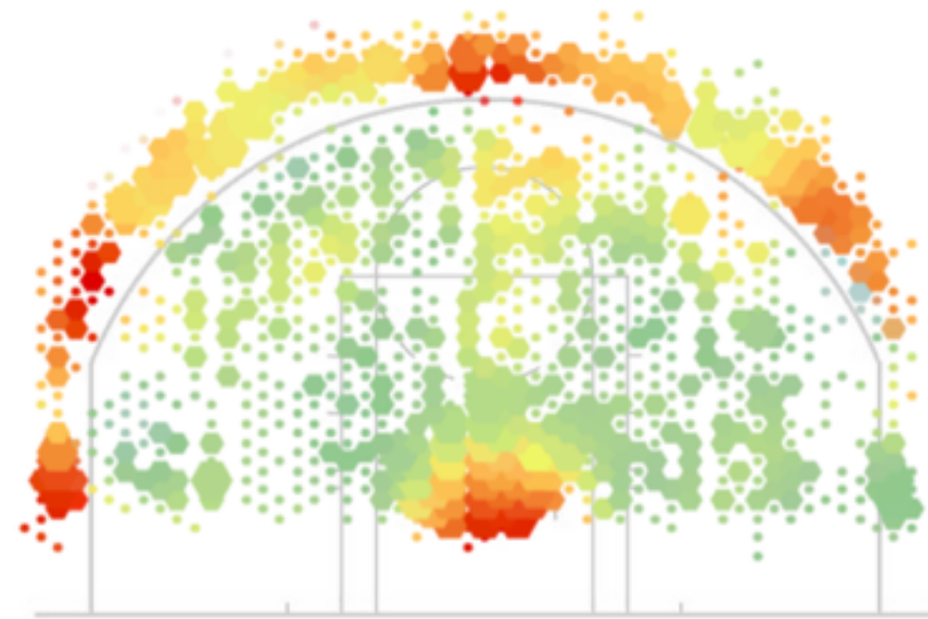
TOTAL SHOTS **5,209** | POINTS PER SHOT **1.01** | F.G. PERCENT **47%**



The Heat rely on player positioning to create isolation plays for LeBron James and Dwyane Wade, often on the left side. The Heat take many fewer 3-point shots than the Thunder.

Oklahoma City Thunder

TOTAL SHOTS **5,228** | POINTS PER SHOT **1.03** | F.G. PERCENT **47.1%**



The Thunder are effective from almost any area on the court and shoot many more 3-point shots than the league average. Kevin Durant and James Harden are potent from the top of the arc.

http://www.nytimes.com/interactive/2012/06/11/sports/basketball/nba-shot-analysis.html

Distance from end line	Distance from LHS	Number of shots	Average number of points	Team
0.25	0.25	5	1.2	Miami Heat
0.25	0.50	10	2.1	Miami Heat
0.25	0.75	5	0.8	Miami Heat
0.25	1.00	3	0.1	Miami Heat

The data explore shots by the Miami Heat and Oklahoma City Thunder basketball teams (over what time period?). For locations on the court, we have the number of attempts from the location, and the average number of points scored from the location.

Each team has its own plot. In each plot, each location is represented by a (hexagonal) point, with a direct mapping between the physical court location and point location (distance from LHS is mapped to horizontal position and distance from end line to vertical position). The color of the point is mapped to the average number of points, and the size to the number of attempts.

The plot is annotated with the usual markings of a basketball court to aid in interpretation of the locations rather than providing axes.

Other simple geometric objects

rectangles

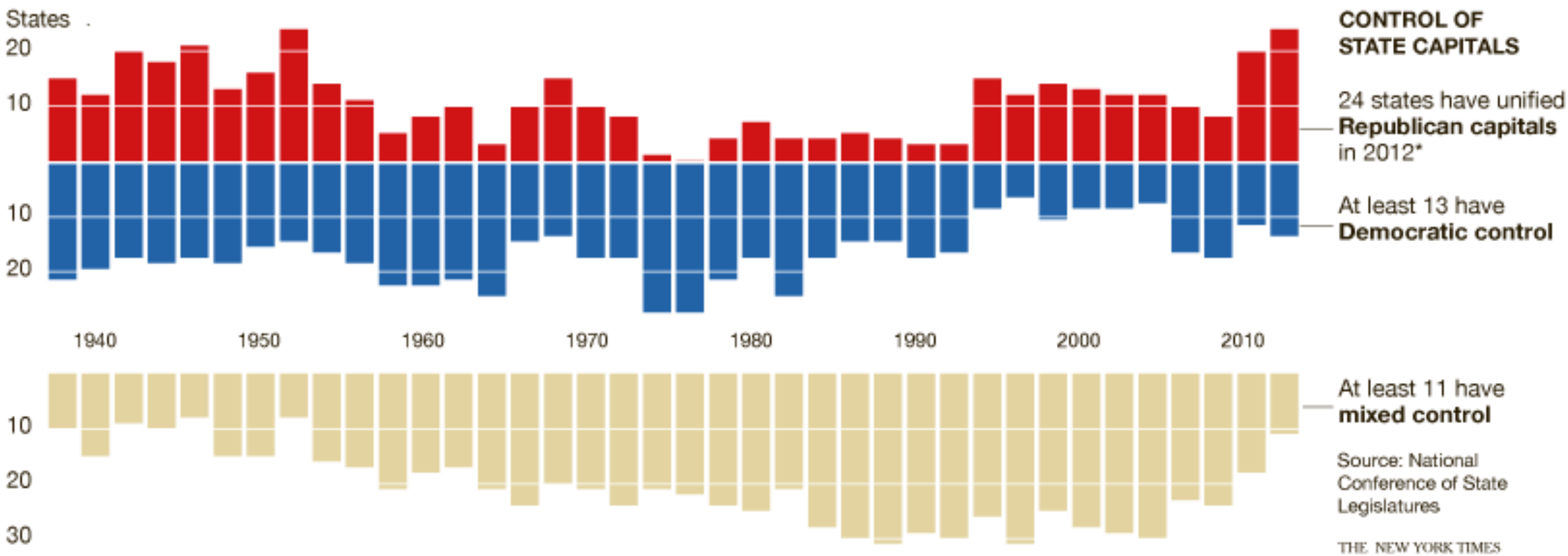
bars (rectangles that have a base at zero)

line segments

Can you guess what their aesthetics should/might be?

Can you describe this?

http://chartsnthings.tumblr.com/post/36904562002/choosing-the-best-form

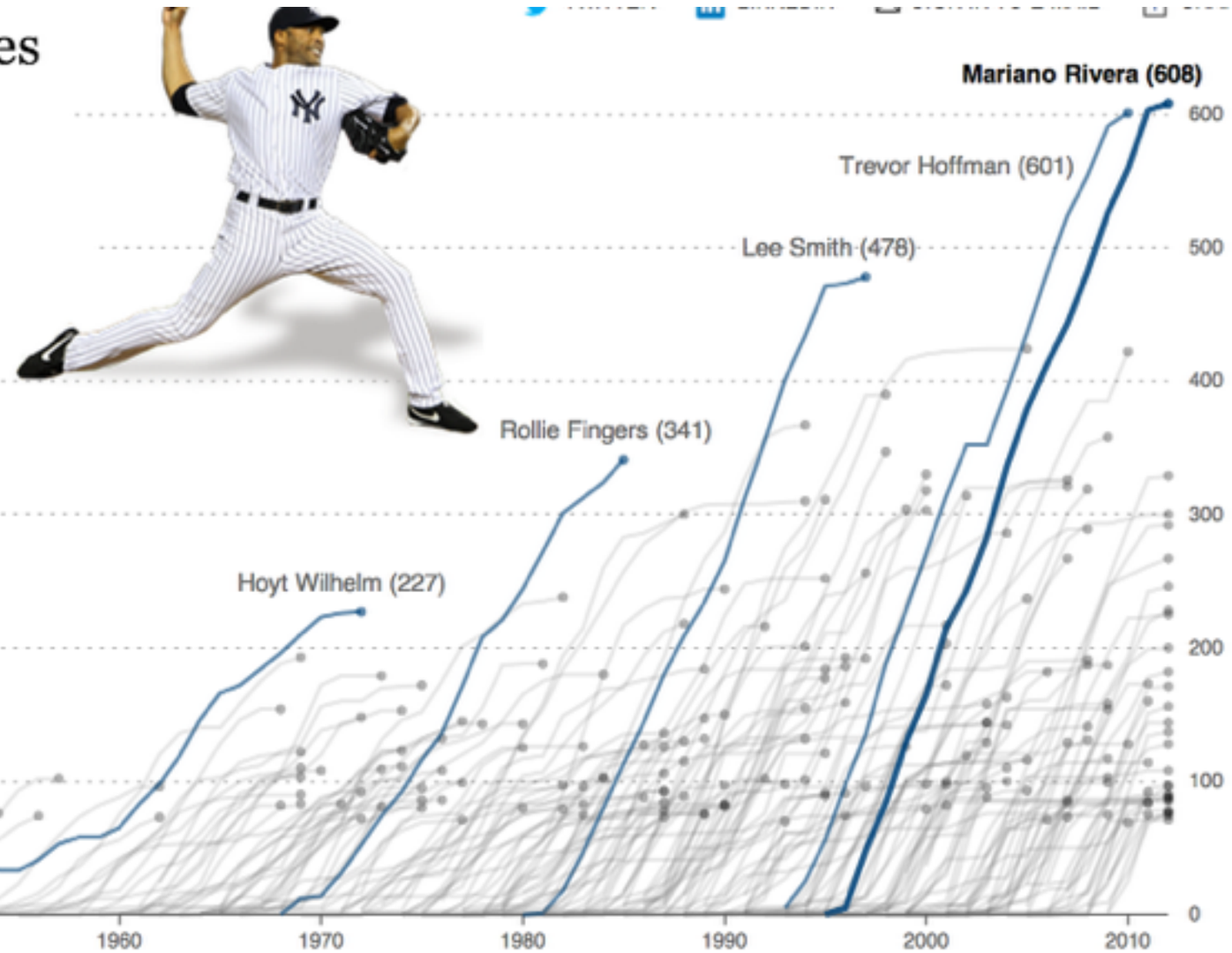


Year	Control	Number of States
1938	Democratic	21
1938	Republican	15
1938	Mixed	10
1940	Democratic	19

Sometimes there is more than one geometric object in a plot…

How Mariano Rivera Compares to Baseball's Best Closers

Mariano Rivera tore his right A.C.L. while shagging fly balls on Thursday, possibly ending his career. Considered the best closer in baseball history, Rivera has more saves than any other pitcher. Below, the cumulative saves of the pitchers with 100 or more.



The closers who broke new hundred-save milestones:

Firpo Marberry (101)
The first reliever to get to 100 cumulative saves, done at a time before relief pitchers were commonplace. (Marberry also started 186 games.)

Hoyt Wilhelm (227)
In addition to being the first pitcher to break the 200 save mark, Wilhelm pitched a no-hitter against the Yankees in 1958.

Rollie Fingers (341)
Known for his handlebar moustache, Fingers was the second relief pitcher inducted into Baseball's Hall of Fame.

Lee Smith (478)
From 1983 to 1995, Smith averaged 35 saves a season, saving no fewer than 25 in any season.

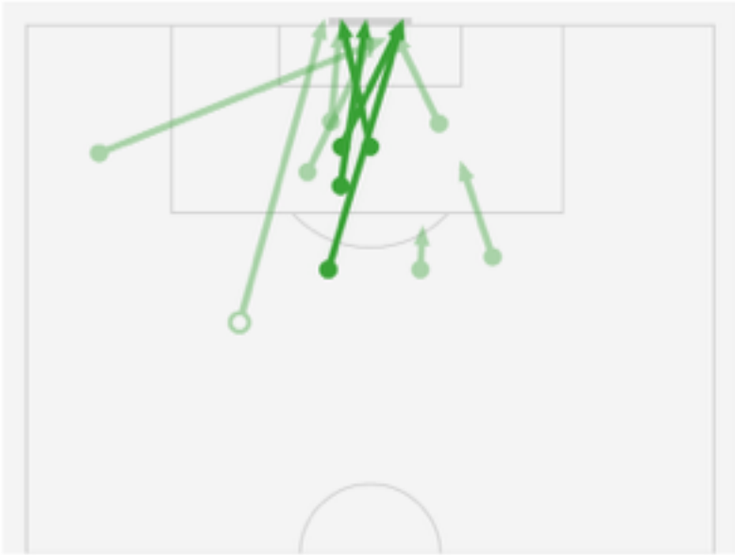
Trevor Hoffman (601)
Hoffman was the first to break the 500 and 600 save marks, despite a 1994 shoulder injury that forced him to change his pitching style.

http://www.nytimes.com/interactive/2012/05/05/sports/baseball/mariano-rivera-and-his-peers.html?ref=baseball

Neymar

Neymar has scored on 36% of the shots he has taken.

Through 3 games

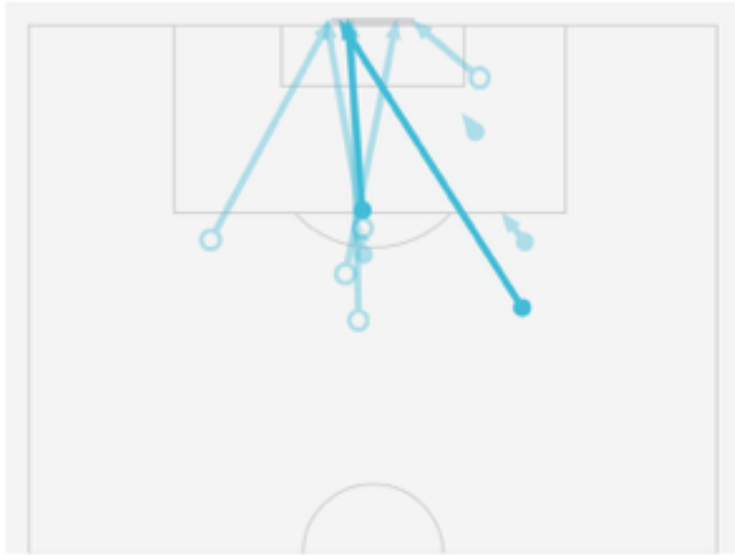


11 shots 10 on target 4 goals

Lionel Messi

Messi saved the day with a late goal against Iran that sent Argentina to the Round of 16.

Through 2 games

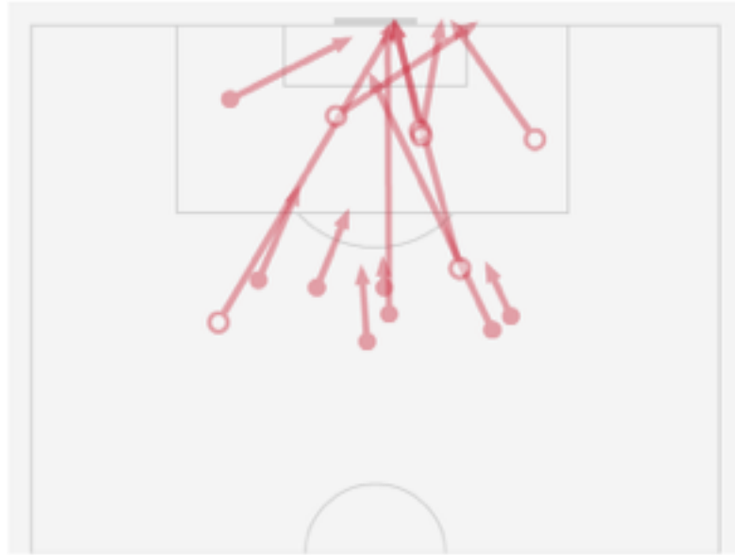


10 shots 5 on target 2 goals

Cristiano Ronaldo

Germany kept Ronaldo from scoring; the United States followed suit, but his deft pass in the final moments led to the tying score.

Through 2 games



14 shots 8 on target 0 goals

http://www.nytimes.com/interactive/2014/06/23/sports/worldcup/brazil-cameroon.html

The ggplot2 grammar

The components of a plot are:

- a default dataset and default set of mappings for variables to aesthetics
- one or more layers each with
  - a geometric object
  - a statistical transformation
  - a position adjustment
  - a dataset
  - a set of aesthetic mappings
- one scale for each aesthetic mapping
- a coordinate system
- a facet specification

Reading

Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3-28.

http://vita.had.co.nz/papers/layered-grammar.pdf

By Monday: at least to section 4

By Friday: whole thing